10-601 Machine Learning, Fall 2009: Homework 1

Due: Wednesday, September 2nd, 10:30 am

Instructions There are 5 questions on this assignment worth the total of 120 points. The last question involves some basic programming. Please hand in a hard copy at the beginning of the class; your code should be printed and attached to the write-up. Refer to the webpage for policies regarding collaboration, due dates, and extensions.

1 Monty's Haunted House [10 pts]

You are in a haunted house and you are stuck in front of three doors. A ghost appears and tells you "Your hope is behind one of these doors. There is only one door that opens to the outside and the two other doors have deadly monsters behind them. You must choose one door."

You choose the first door. The ghost tells you "Wait! I will give you some more information." The ghost opens the second door and shows you that there was a horrible monster behind it, then asks you "Would you like to change your mind and take the third door instead?"

Which strategy is better: to stick with the first door, or to change to the third door? For each of the following possibilities, determine probabilities that the exit is behind the first and the third door, given that the ghost opened the second door.

1. The ghost uses the same strategy as discussed in class. He always opens a door you have not picked with a monster behind it. If both of the unopened doors hide monsters, he picks each of them with equal probability. [2 pts]

2. The ghost has a slightly different strategy. If both of the unopened doors hide monsters, he always picks the second door. [4 pts]

3. Finally, suppose that if both of the unopened doors hide monsters, the ghost always picks the third door. [4 pts]

2 Medical Testing [15 pts]

There is a disease which affects 1 in 500 people. A \$100.00 dollar blood test can help reveal whether a person has the disease. A positive outcome indicates that the person *may* have the disease. The test has perfect sensitivity (true positive rate), i.e., a person who has the disease tests positive 100% of the time. However, the test has 99% specificity (true negative rate), i.e., a healthy person tests positive 1% of the time.

1. A randomly selected individual is tested and the result is positive. What is the probability of the individual having the disease? [5 pts]

2. There is a second more expensive test which costs \$10,000.00 dollars but is exact with 100% sensitivity and specificity. If we require all people who test positive with the less expensive test to be tested with the more expensive test, what is the expected cost to check whether an individual has the disease? [5 pts]

3. A pharmaceutical company is attempting to decrease the cost of the second (perfect) test. How much would it have to make the second test cost, so that the first test is no longer needed? That is, at what cost is it cheaper simply to use the perfect test alone, instead of screening with the cheaper test as described in part 2? [5 pts]

3 Products of Expectations [20 pts]

We showed in class that $\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y]$ if X and Y are independent.

1. Prove that the converse is also true when X and Y are binary, i.e., if X and Y take values in $\{0, 1\}$ and $\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y]$, then X and Y are independent. [10 pts]
2. Does the converse hold if X is required to take values in $\{0, 1\}$, but Y can be arbitrary? Prove or give a counterexample. [10 pts]

4 Balls and Bins [15 pts]

Suppose we have n bins and m balls. We throw balls into bins independently at random, so that each ball is equally likely to fall into any of the bins.

1. What is the probability of the first ball falling into the first bin? [2 pts]
2. What is the expected number of balls in the first bin? [3 pts]

   Hint 1: Define an indicator random variable representing whether the i-th ball fell into the first bin:

   $$X_i = \begin{cases} 1 & \text{if } i\text{-th ball fell into the first bin} \\ 0 & \text{otherwise.} \end{cases}$$

   Hint 2: Use linearity of expectation.
3. What is the probability that the first bin is empty? [5 pts]
4. What is the expected number of empty bins? [5 pts]

   Hint 3: Define an indicator for the event "bin j is empty" and use linearity of expectations.

5 Tacky Programming [60 points]

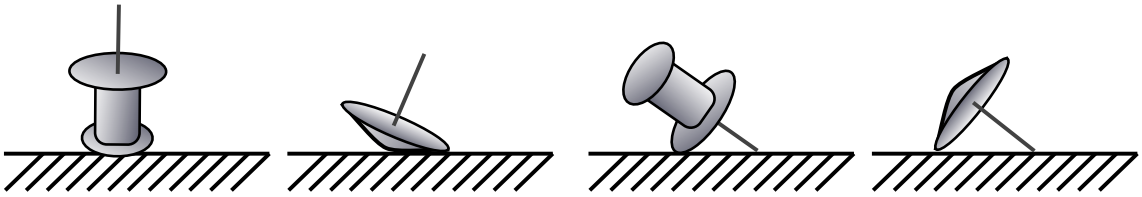
In this problem you will have an opportunity to become familiar with your numerical computing environment. We highly recommend that you use Matlab and will only provide help with Matlab code. You may obtain Matlab from `http://www.cmu.edu/computing/software/all/matlab/index.html` or by using many of the facilitized systems on campus. For this problem you will need to download the data from `http://www.cs.cmu.edu/~ggordon/10601/hws/hw1/hw1data.zip`. For all plots, please include titles and axis labels (see Matlab commands `xlabel`, `ylabel`, and `title`). Please include all code with your homework submission.

Problem Setup: A fun game, among graduate students, is tossing thumbtacks[1] to see if they land facing up or facing down as seen in Fig. 1. We can treat the orientation of a thumbtack as random variable Y which can either be *up* or *down*. To make the game more interesting, a friend of mine commissioned a multinational team of unbiased experimental physicists to build the perfect set of 100 uniquely tuned thumbtacks. For each thumbtack $X \in \{1, \ldots, 100\}$ the probability that the thumbtack will land facing up is given by:

$$\mathbf{P}(Y = \text{up} \mid X = x) = \frac{x}{100} \tag{5.1}$$

[1] Careful! Thumbtacks are believed to be sharp.

Hence, thumbtack 1 lands facing up with probability 0.01 and thumbtack 100 lands facing up with probability 1. After cleaning out my office, I filled three jars, $J \in \{1, 2, 3\}$, with varying amounts of each type of thumbtack. I then thoroughly and meticulously shook each jar. You can find the counts of each type of thumbtack for all three jars in `jars.csv`. The first column is the thumbtack id X, and the remaining 3 columns contain the counts for each jar.



(a) Up Thumbtacks (b) Down Thumbtacks

Figure 1: Thumbtacks landing up and down

5.1 Looking at the Data [10 pts]

To familiarize yourself with the problem use the `bar` function in Matlab to plot:

1. the *probability* $Y = \text{up}$ for each type of thumbtack
2. the *probability* of drawing each type of thumbtack from jar 1
3. the *probability* of drawing each type of thumbtack from jar 2
4. the *probability* of drawing each type of thumbtack from jar 3

To save paper, please consider plotting several plots on the same page (see `help subplot` for how to do this in Matlab).

5.2 The Likelihood of Multiple Observations [5 pts]

Using the `bar` function in Matlab, plot the probability of obtaining the sequence (u, u, d, u, d) of random tosses for each type of thumbtack.

5.3 Writing Bayes Rule [5 pts]

Suppose I give you $\mathbf{P}(Y = \text{up} \mid X)$ and $\mathbf{P}(X)$. Write down equations I would use to compute $\mathbf{P}(X \mid Y = \text{down})$ and $\mathbf{P}(X \mid Y = \text{up})$.

5.4 Posterior Dependence on Number of Observations [20 pts]

Suppose I randomly select a thumbtack from the first jar $J = 1$. For each of the following scenarios, plot the posterior distribution over thumbtack types given the outcomes of the tosses in that scenario

1. I toss the thumbtack only once and it lands facing down.
2. I toss the thumbtack 3 times and each time it lands facing down.
3. I toss the thumbtack 5 times and 3 times it lands facing up and 2 times it lands facing down.
4. I toss the thumbtack 40 times and 25 times it lands facing up and 15 times it lands facing down.

Based on the last (largest) set of tosses, which thumbtack type X do you think I selected? How does increasing the number of observations affect the posterior distribution? How does the change in posterior distribution relate to your confidence in the prediction of which thumbtack type was used?

5.5 Marginalizing The Unobserved [20 pts]

Suppose I randomly select a jar with probability $\mathbf{P}(J = 1) = 0.25$, $\mathbf{P}(J = 2) = 0.25$, and $\mathbf{P}(J = 3) = 0.50$. Then from the chosen jar, I randomly draw a thumbtack. I then toss the thumbtack 10 times and it lands facing up 8 times and facing down 2 times.

1. Plot the posterior distribution over thumbtack types given the outcome of the tosses.
2. Plot the posterior distribution over the jars given the outcome of the tosses.
3. Which thumbtack type attains the largest posterior? Which jar attains the largest posterior?